statstutor

community project

encouraging academics to share statistics support resources

Statistical Methods 11. Correlation and Simple Linear Regression

Based on materials provided by Coventry University and Loughborough University under a National HE STEM Programme Practice Transfer Adopters grant

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Workshop outline

We will consider:

- Correlation coefficients:
  - Pearson's correlation
  - Spearman's rank correlation
- Simple linear regression
- The importance of outliers and residuals

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Correlation

- Correlation is a **measure** of the **strength** of the **linear** association between two **scale** variables
- Correlation is often used inappropriately when "association" is meant
- The correct terminology is "**correlation coefficient**"
- We usually calculate at **Pearson's** correlation coefficient

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Pearson's correlation

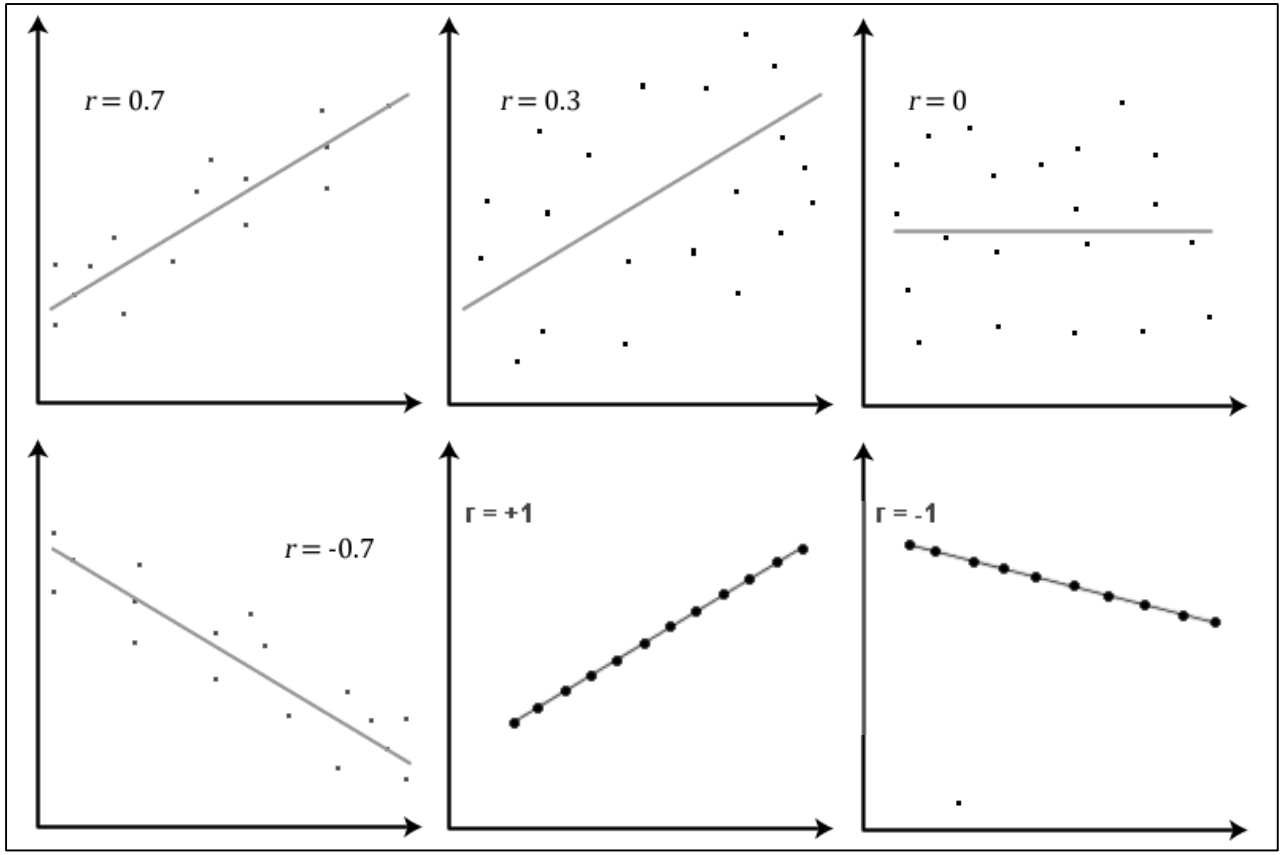
- Symbols used for the correlation coefficient
  - ρ ('rho') is used for the population value
  - r is used for its estimate
- ρ and r are always between -1 and +1
- Positive ρ or r implies y ↑ as x ↑
- Negative ρ or r implies y ↓ as x ↑
- The correlation coefficient is sensitive to outliers. Look at scatter plot using the 'Graphs' menu in SPSS

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Correlation coefficient: examples

- The **sign** of *r* depends on the **direction** of the fitted line (**not** its slope)
- The **magnitude** of *r* depends on how closely the data points are **aligned** (all on a line means the coefficient is +1 or -1)



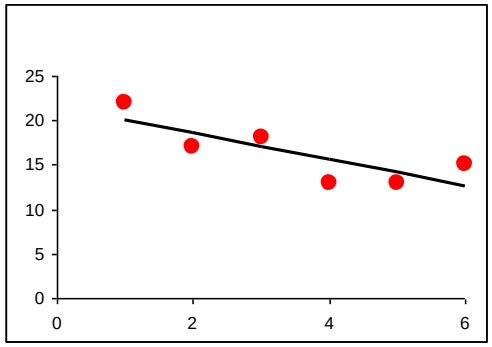
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

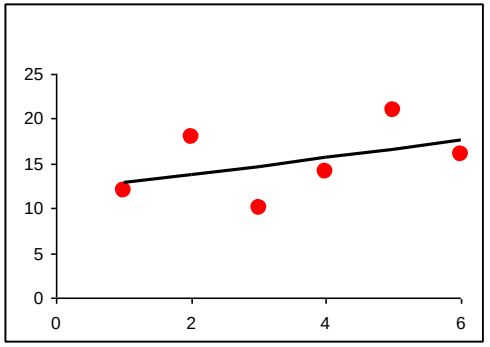
Activity

Estimate *r* and describe the correlation in each diagram

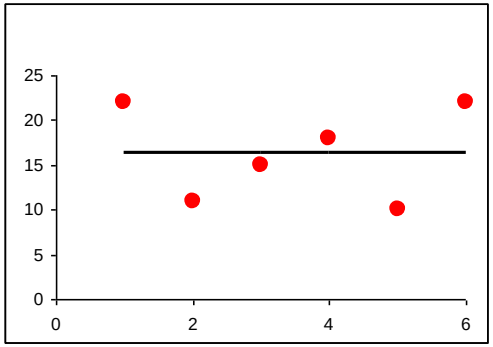
Strong negative correlation



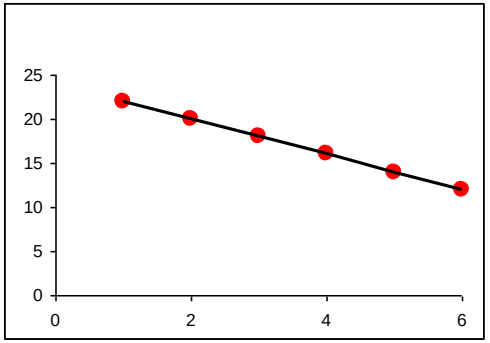
Weak positive correlation



No correlation



Perfect negative correlation



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

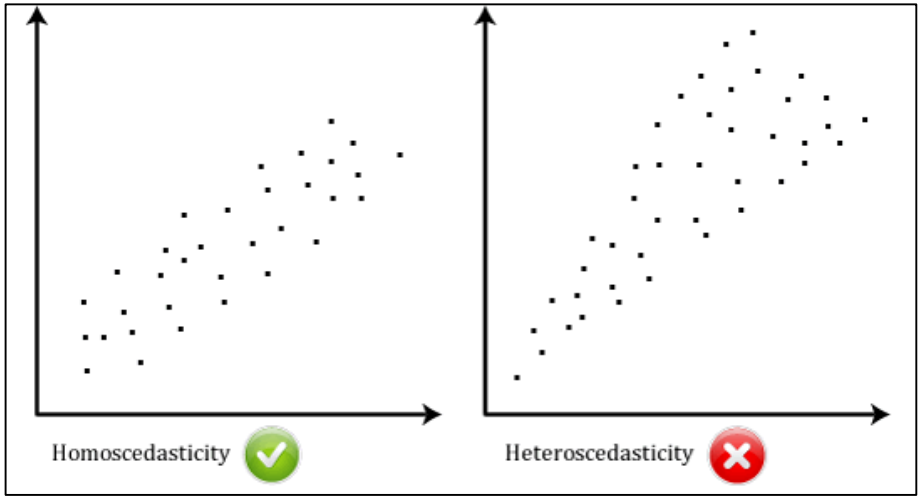
Interpretation of *r*

r	**Interpretation**
0.1	Small
0.3	Medium
0.5	Large

Source: Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Assumptions

1. Both variables are normally distributed
2. The relationship between the two variables is **linear**.
3. The relationship between the two variables is **homoscedastic** (i.e. the variance of one variable is the same for all the values of the other variable). We can test 2 and 3 by looking at the scatter plot and observing whether the data points form a "roughly symmetrical, cigar-shaped pattern" about the regression line.



If these assumptions or robust exceptions (see later) are not met, we should use Spearman's rank correlation (see later).

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 1: Forest measurement

A study of the relationship between basal growth and crown volume of 62 trees is reported by Avery and Burkhart (1994):

- **Basal Growth** is the change in cross sectional area (in square feet) at chest height in one year
- **Crown Volume** is the increase in the total volume (in cubic feet) of the tree above the first branch

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Set up the data file

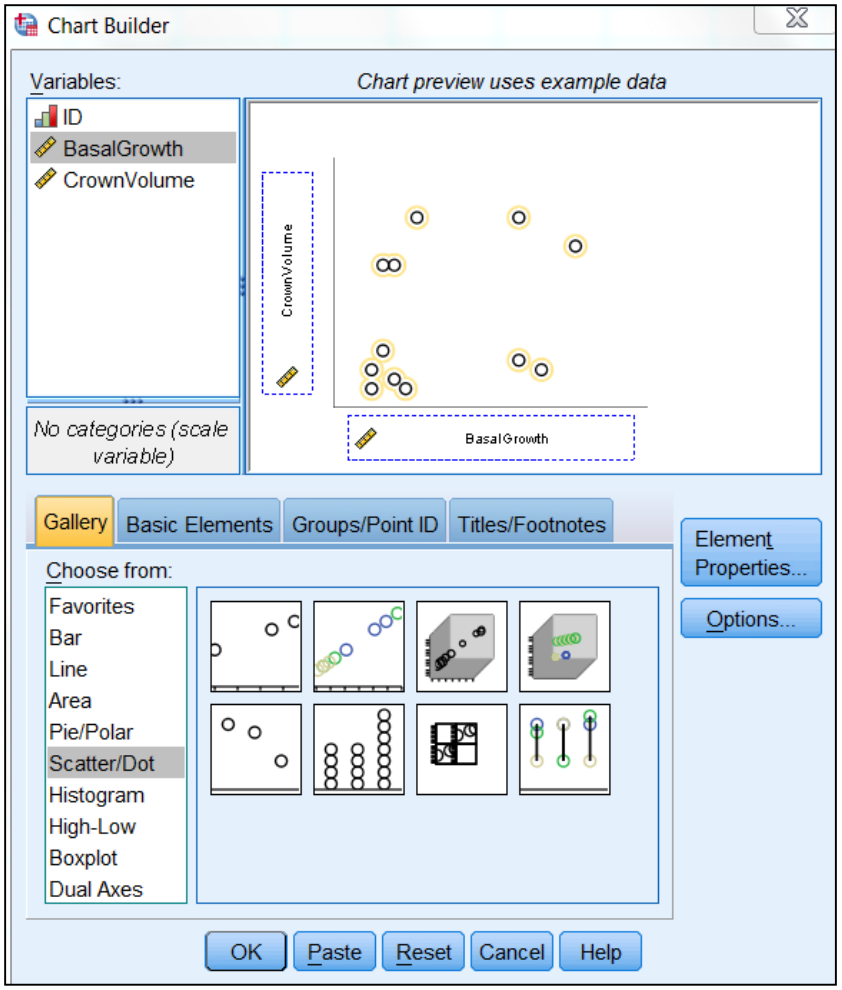
- Open the Excel file BasalCrownData.xlsx from associated with this presentation
- Copy the data into a data window SPSS
- Set up the variable names as in the Excel file
- Set the measures to ordinal, scale and scale respectively
- Set the number of decimal places to 0, 2 and 0 respectively
- Save the file as BasalCrown.sav

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Create a scatter plot

- Select Graphs – Chart Builder
- Select Scatter/Dot from the Choose from: menu on the Gallery tab
- Click and drag the first scatter plot into the chart preview area
- Click and drag *BasalGrowth* onto the x-axis and *CrownVolume* onto the y-axis

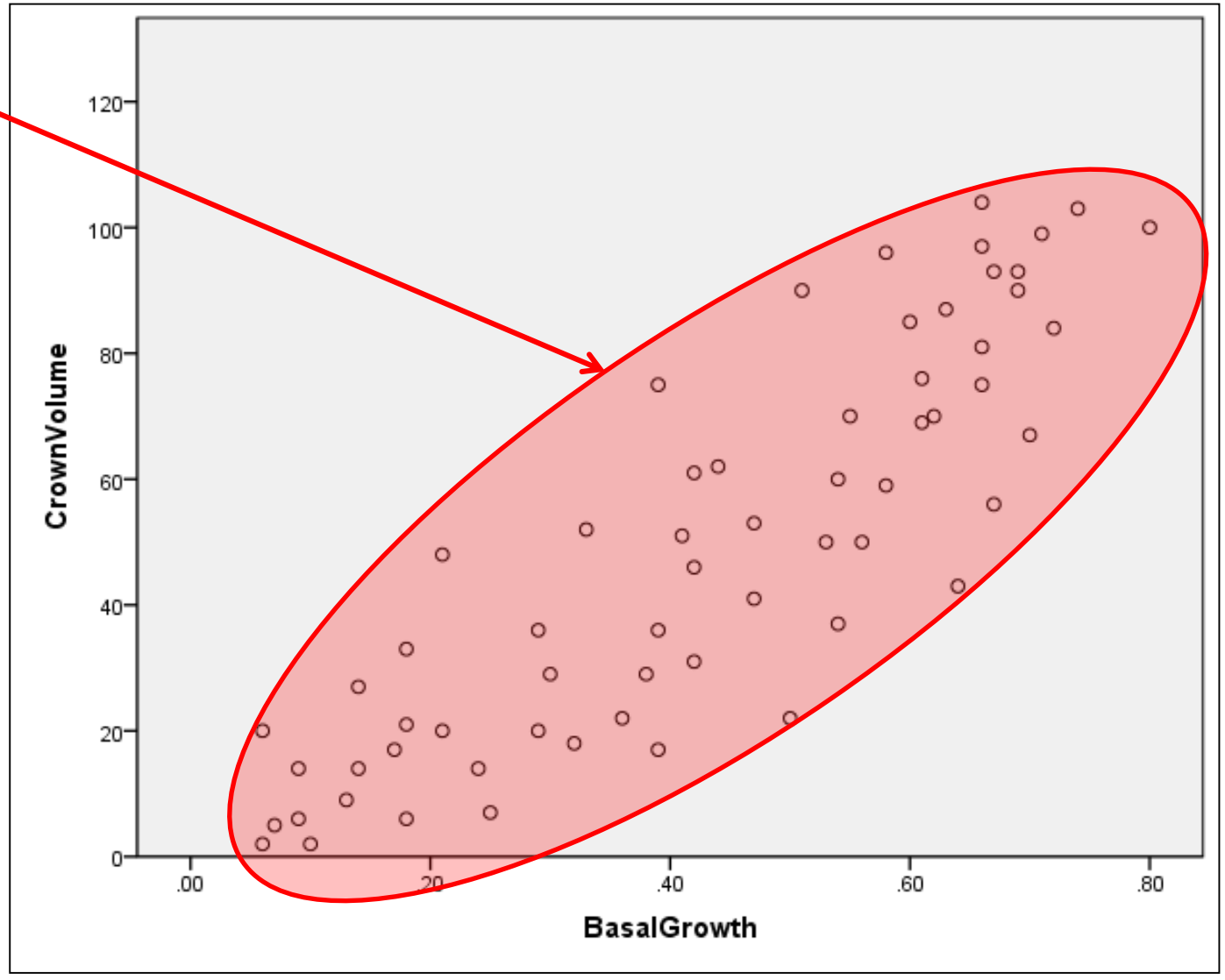


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Dispersion of points is 'cigar shaped'

The data values appear to meet the assumptions for a correlation analysis

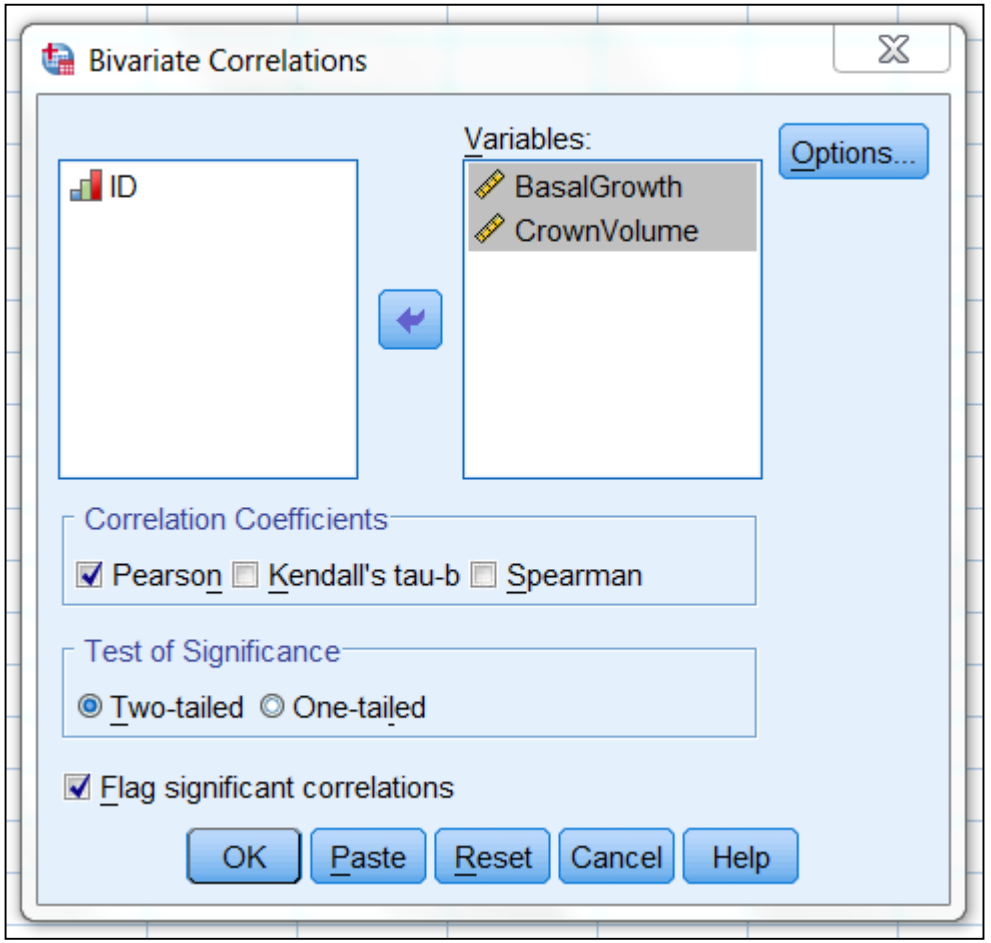


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Carry out the correlation analysis

- Select Analyze > Correlate > Bivariate
- Select the two variables
- The default correlation analysis is Pearson
- The default significance is 2-tailed



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Returns a correlation coefficient *r* of 0.871
- Analysis is repeated because it is comparing each variable on the list with each other in turn – generally look at the cells below the leading diagonal when there are more than 2 variables

Correlations

		BasalGrowth	CrownVolume
BasalGrowth	Pearson Correlation	1	.871**
	Sig. (2-tailed)		.000
	N	62	62
CrownVolume	Pearson Correlation	.871**	1
	Sig. (2-tailed)	.000	
	N	62	62

**. Correlation is significant at the 0.01 level (2-tailed).

Significance level is actually 0.001, **not** 0.01, as indicated by the footnote, because the p-value of '0.000' actually means < 0.0005, so it is clearly also < 0.001

The null hypothesis is $\rho = 0$

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Correlation caveats (1)

- **Correlation does not imply causation**:
  - Both x and y could be influenced by z
  - E.g. there is a positive correlation between wearing a waistcoat watch (x) and heart disease (y) because of the influence of wealth and diet (z)
- With large samples even small correlation coefficients can be statistically significant:
  - Think about what would be of practical importance
- Beware of outliers!
  - Correlation coefficients are sensitive to outliers
  - However, outliers should never be removed without a valid reason being given

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Correlation caveats (2)

$r = 0$ indicates no linear association:

- Low absolute values of r do not necessarily mean that the variables are not related – any relationship may be non-linear
- r^2 (a.k.a. R^2) indicates the amount of variability 'explained' by the relationship between the two variables
- $r = 0.7$, gives $r^2 = 0.49$, i.e. only 49% of the variability is explained by the relationship
- Different absolute values of r are meaningful in different contexts

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Robustness exceptions

- Correlation calculations are not robust to violations of homo-scedasticity – the data could be transformed in this case
- The hypothesis test for $\rho = 0$ is robust to extreme violations of normality. However Spearman's rank correlation (see later) is sometimes a more powerful test.
- Interpretations of the value of *r* can be completely meaningless if the joint distribution of the two variables is too different from a binormal distribution

References:

Asuero, A. Sayago, A. & González, A. (2006) The Correlation Coefficient: An Overview, *Critical Reviews in Analytical Chemistry*, 36: 41–59

Fowler, R. L. (1987) Power and robustness in product-moment correlation, *Applied Psychological Measurement ,* 11(4): 419-428

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 2: Advertising cds

A record company decided to advertise 200 different cds and measure the sales of the cds the week after they were advertised (in thousands) against the amount spent on advertising (thousands of pounds).

Source: (Field, 2013: Section 8.3)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

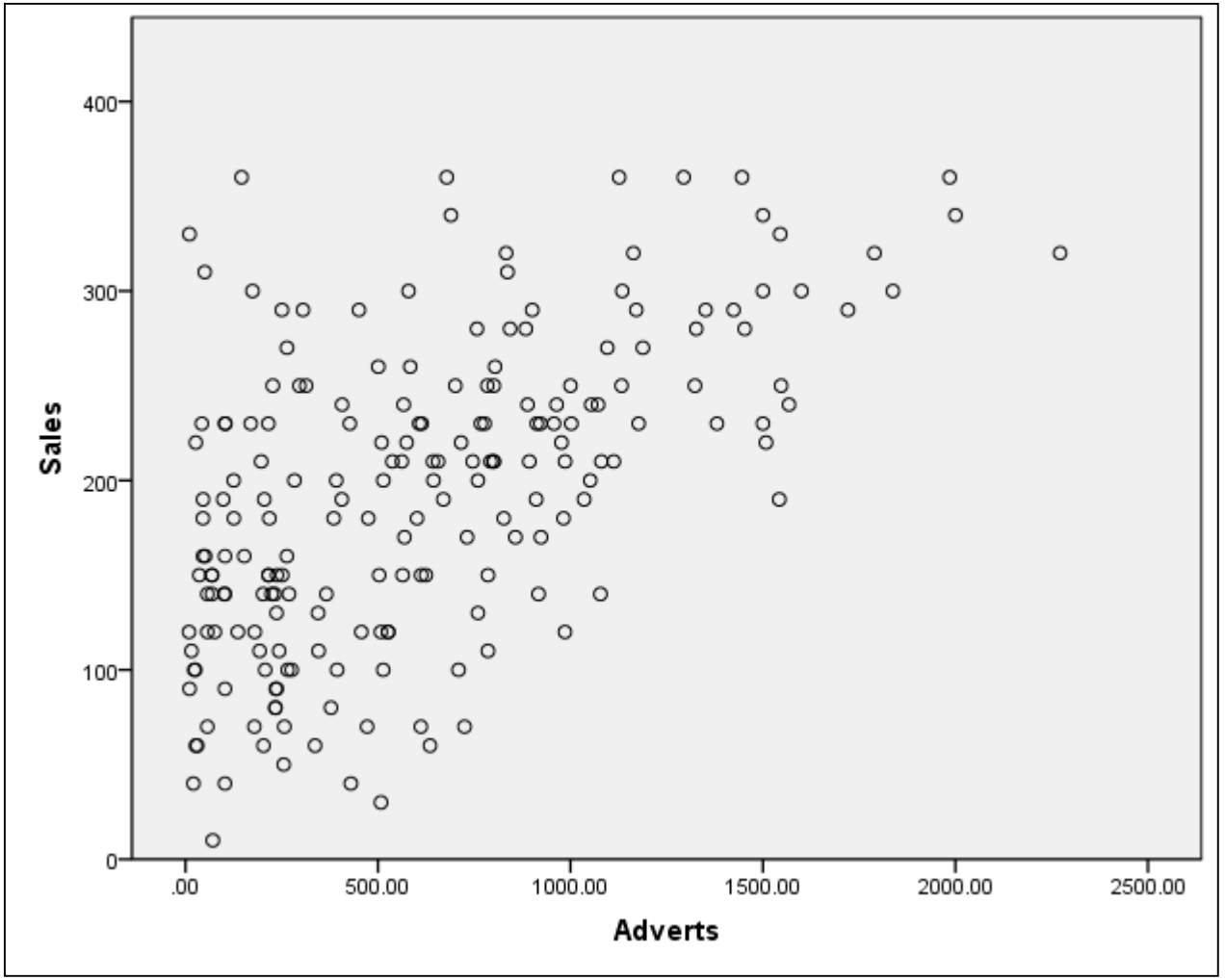
Set up the data file

- Open the Excel file CDSalesData.xlsx associated with this presentation
- Copy the data into a data window SPSS
- Set up the variable names
- Set the measures to scale and scale
- Set the number of decimal places to 2 and 0
- Save the file as CDSales.sav
- Create a scatter plot of *Sales* against *Adverts*

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Distribution is clearly heterosc-edastic
- But variables are clearly associated



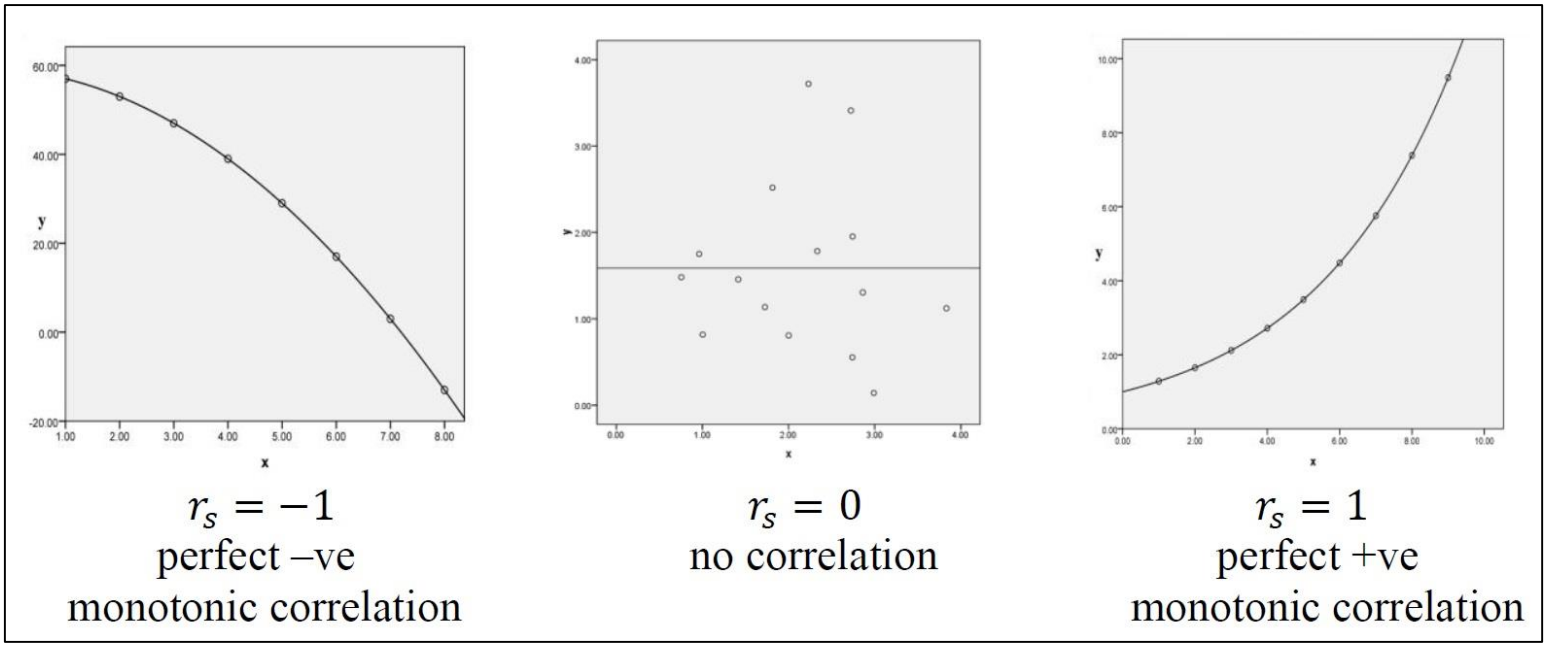
$\Rightarrow$ Need to use Spearman's rank correlation

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Spearman's rank correlation

- Similar to Pearson's rank but can be applied to ordinal data as well as scale data
- Measures how consistently one variable increases or decreases as a second variable increases (monotonic)
- Represented by the symbol r_s



$r_s = -1$
perfect –ve
monotonic correlation

$r_s = 0$
no correlation

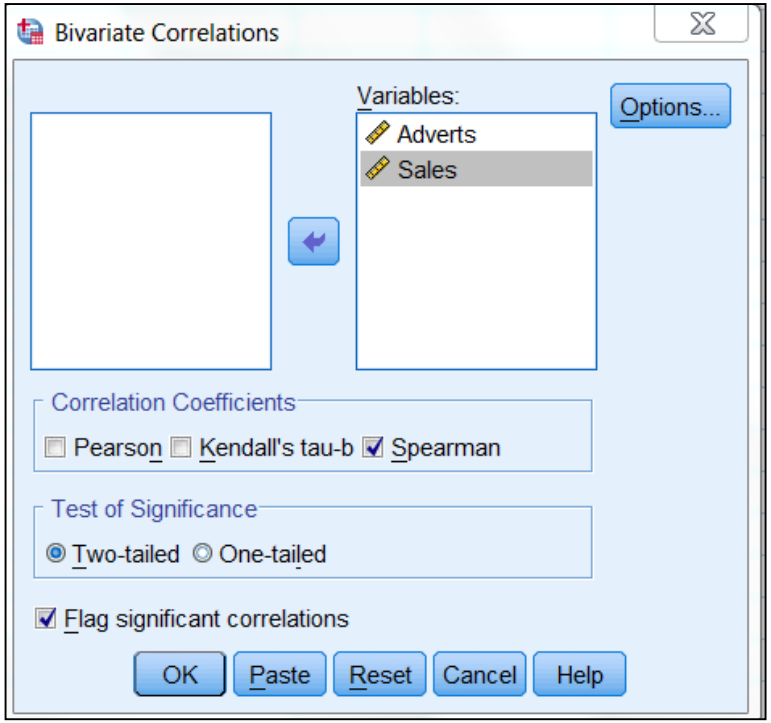
$r_s = 1$
perfect +ve
monotonic correlation

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Spearman's rank in SPSS

- Select *Analyze – Correlate – Bivariate…*
- Select both variables
- Select Spearman for the correlation coefficients



Correlations

			Adverts	Sales
Spearman's rho	Adverts	Correlation Coefficient	1.000	.554**
		Sig. (2-tailed)	.	.000
		N	200	200
	Sales	Correlation Coefficient	.554**	1.000
		Sig. (2-tailed)	.000	.
		N	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

Returns a Spearman correlation coefficient of 0.554

Probability < 0.0005 so the association is significant at 0.001 level

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Regression analysis

- Uses data to build a **statistical model** to **describe the relationship** between different quantities or variables
- **Simple linear regression** describes a linear relationships between two scale (continuous) variables:
  - The x variable is the **independent**, or **predictor**, variable
  - The y variable is the **dependent**, or **outcome**, variable

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Simple linear regression

The model is: $y = b_0 + b_1x$

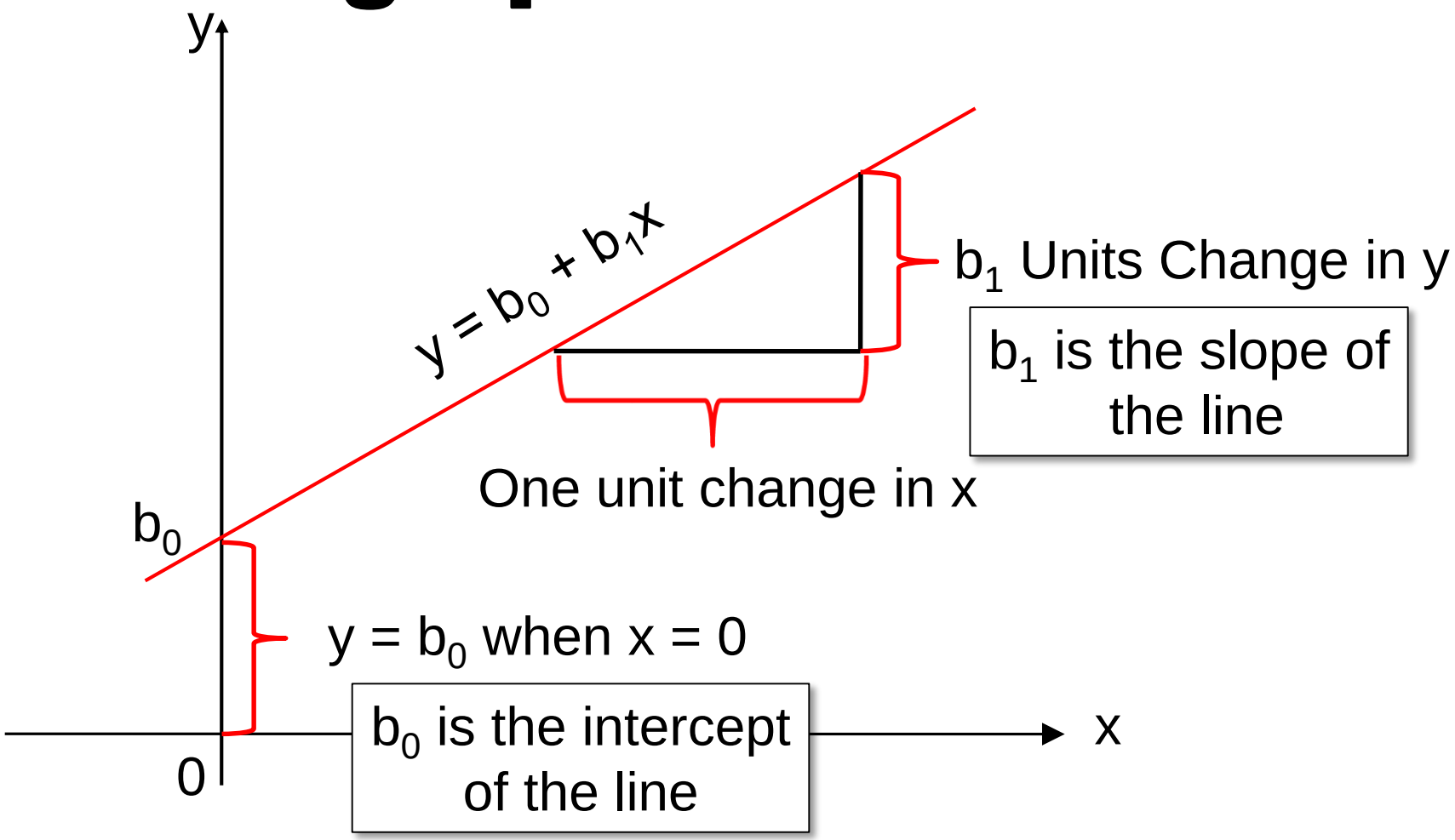
Where:

- b_0 is the **intercept** or constant term
- b_1 is the **slope** of the line
    - b_0 and b_1 are known as **coefficients**
- (You may be more familiar with the notation $y = a + bx$ or $y = mx + c$)
- Linear regression fits the 'best' straight line to the data
- There are different ways of defining 'best'
- The most common method is called **least squares**

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Simple linear regression – graphical view



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Residuals

- For “real” data sets there will always be a difference between what we observe and what our model predicts
- We adjust for this difference by adding an error term in the model:

$$y = b_0 + b_1x + e$$

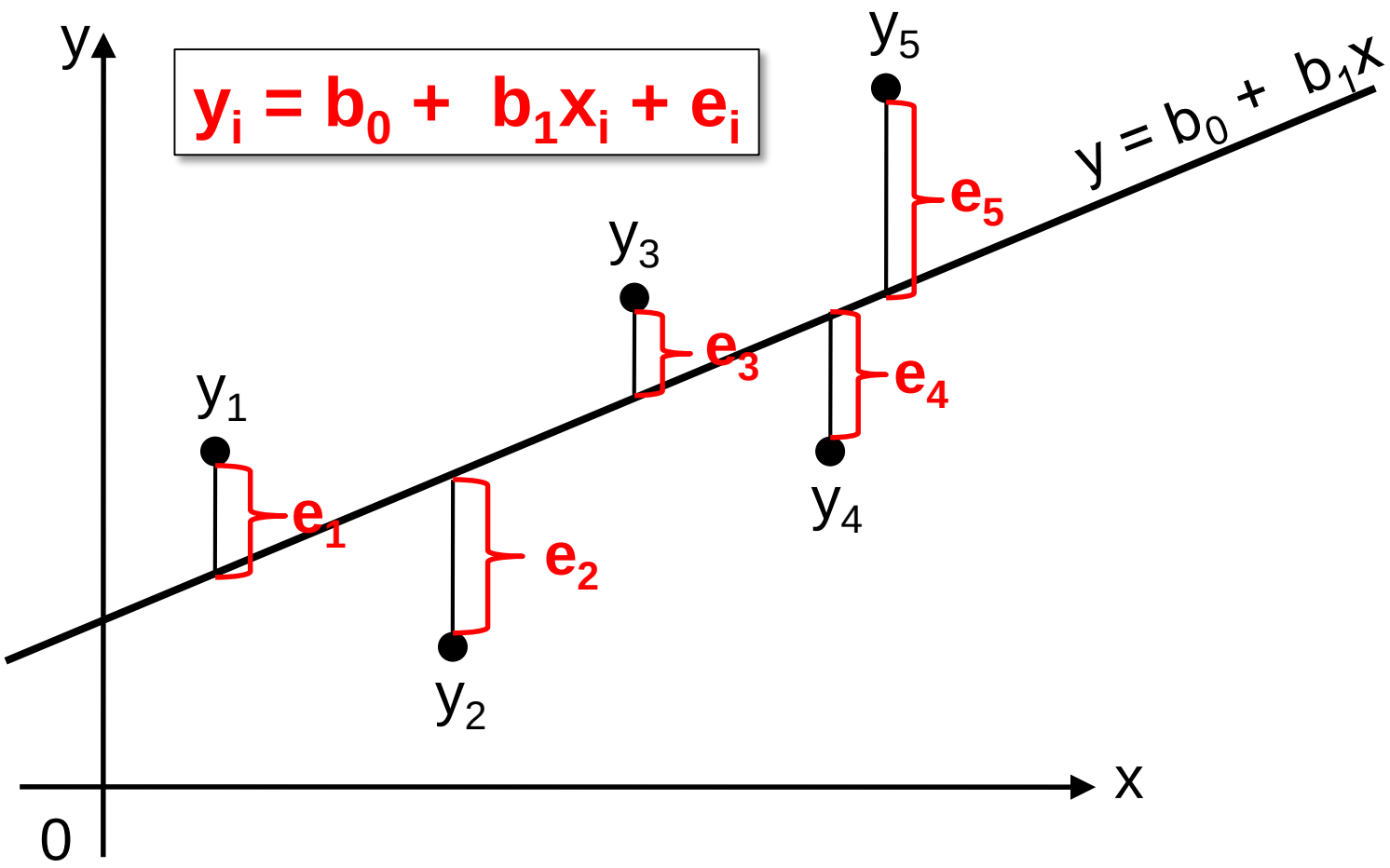
- **Residuals** are then defined as:

$$\begin{aligned} e &= y - (b_0 + b_1x) \\ &= \text{observed value - predicted value} \\ &= \textbf{error} \end{aligned}$$

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Fitting a regression line



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Least squares linear regression

- Regression analysis fits a line through the data using the method of least squares
- The least squares method minimises the sum of squares of the vertical distances of the observed data from the fitted line
- I.e. least squares minimizes the sum of the squared residuals
- This is like drawing squares next to each residual and minimising the sum of the area of these squares

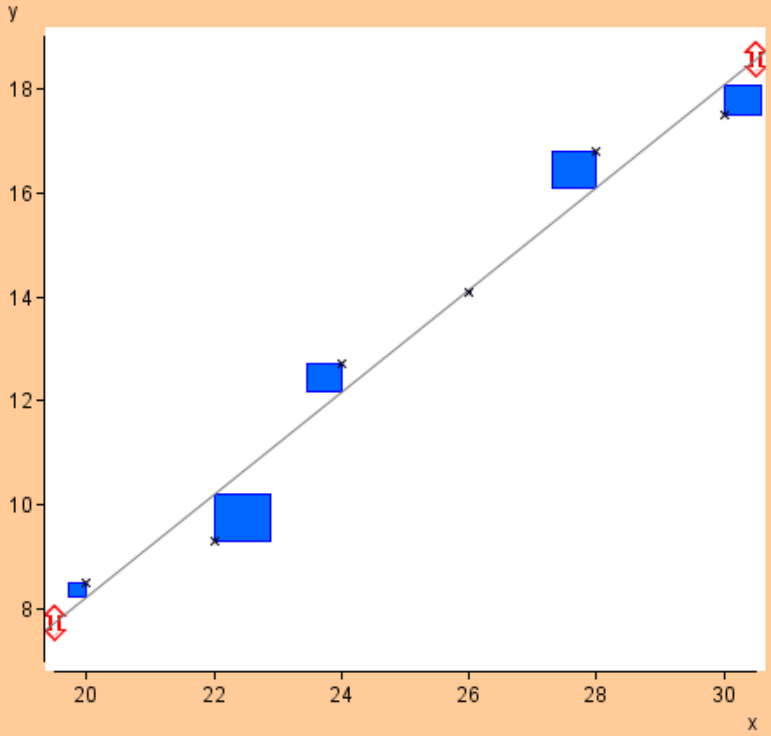
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Least squares linear regression

From http://cast.massey.ac.nz/core/index.html?book=general Section 3.4.4

In the diagram below, the squared residuals are represented by the areas of the boxes at each data point. **Drag** the red arrows to position the line in such a way that the residual sum of squares (the total area of the boxes) is as small as you can achieve.



After trying by hand, click the button **Least squares** to let the computer determine the best values.

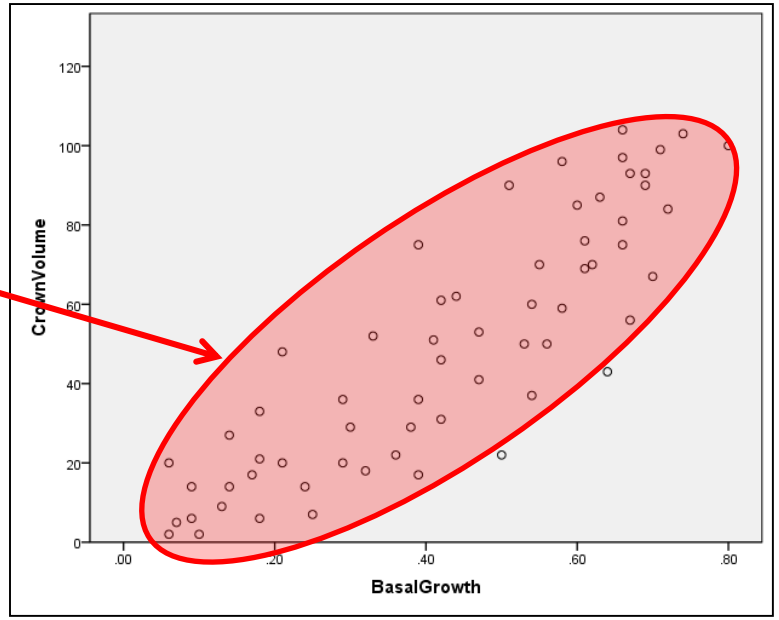
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Model assumptions

1. The mean response (y - e) varies linearly with predictor (x)
2. The unexplained variation (e) is normally and independently distributed with constant variance (i.e. independent of x, or it is homoscedastic)

These are both shown by a ‘cigar shaped’ scatter plot around a straight line



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Activity

- Fit a least squares linear regression model to the BasalCrown data set:

$$CrownVolume = b_0 + b_1 \times BasalGrowth$$

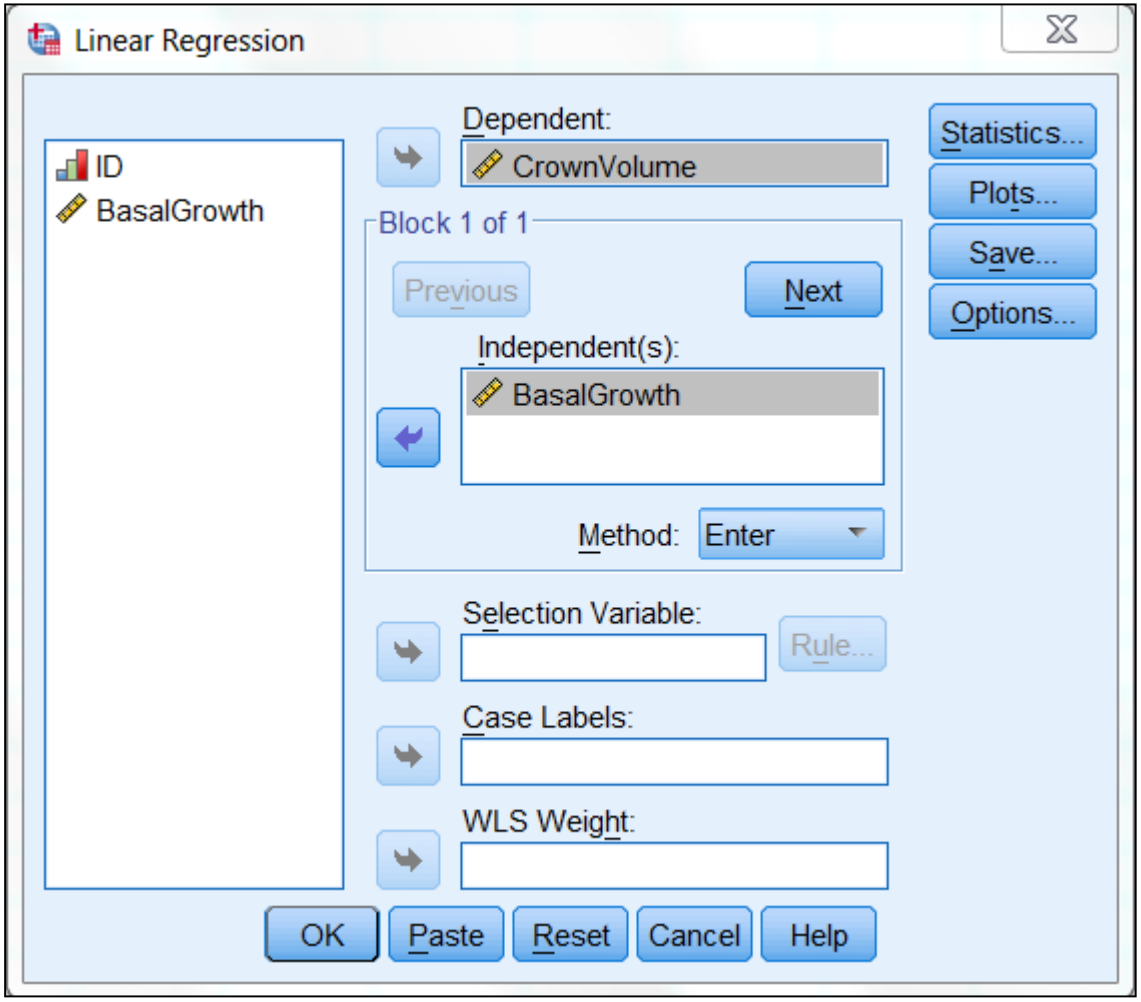
- Check the model for significance
- What are the estimated values of b_0 and b_1?
- Write out the model

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

With the file CrownBasal.sav:

- Select Analyze > Regression > Linear
- Choose *CrownVolume* as the dependent variable and *BasalGrowth* as the independent variable



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5.452	4.435		-1.229	.224
	BasalGrowth	127.273	9.263	.871	13.741	.000

a. Dependent Variable: CrownVolume

$b_0 = -5.452$

$b_1 = 127.273$

The constant coefficient (b_0) estimate is not significant (H_0: $b_0 = 0$) – this is normally ignored

The *BasalGrowth* coefficient (b_1) estimate is highly significant ($p < 0.001$) (H_0: $b_1 = 0$)

Therefore the model is:

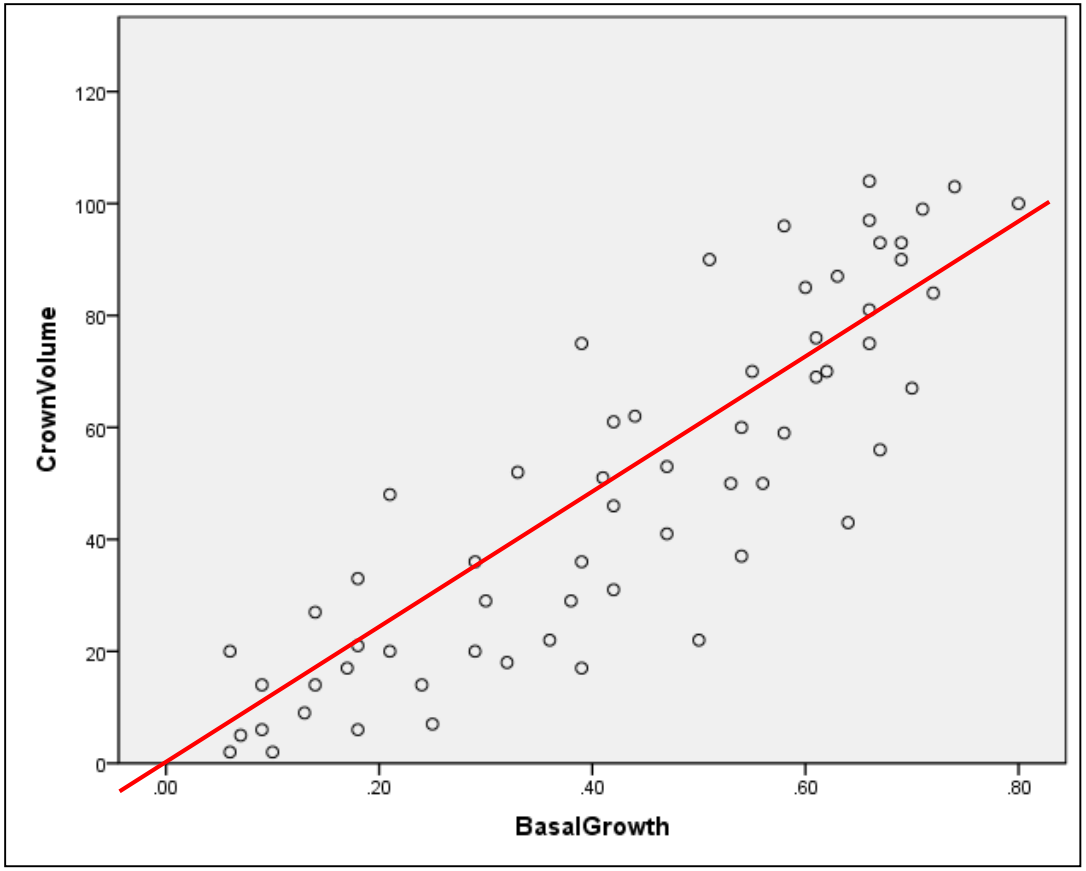
CrownVolume = -5.452 + 127.273×*BasalGrowth*

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

The fitted model

$$CrownVolume = -5.452 + 127.273 \times BasalGrowth$$

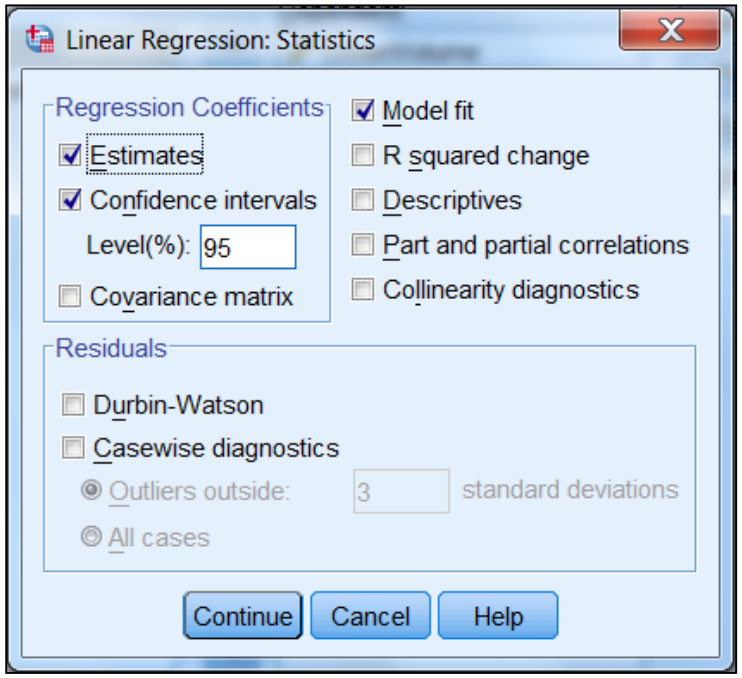


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Confidence intervals for b_0 and b_1

- Redo the analysis
- Select Statistics…
- Select Confidence intervals
- Gives upper and lower bounds for the confidence intervals for the two coefficients



Coefficients[a]

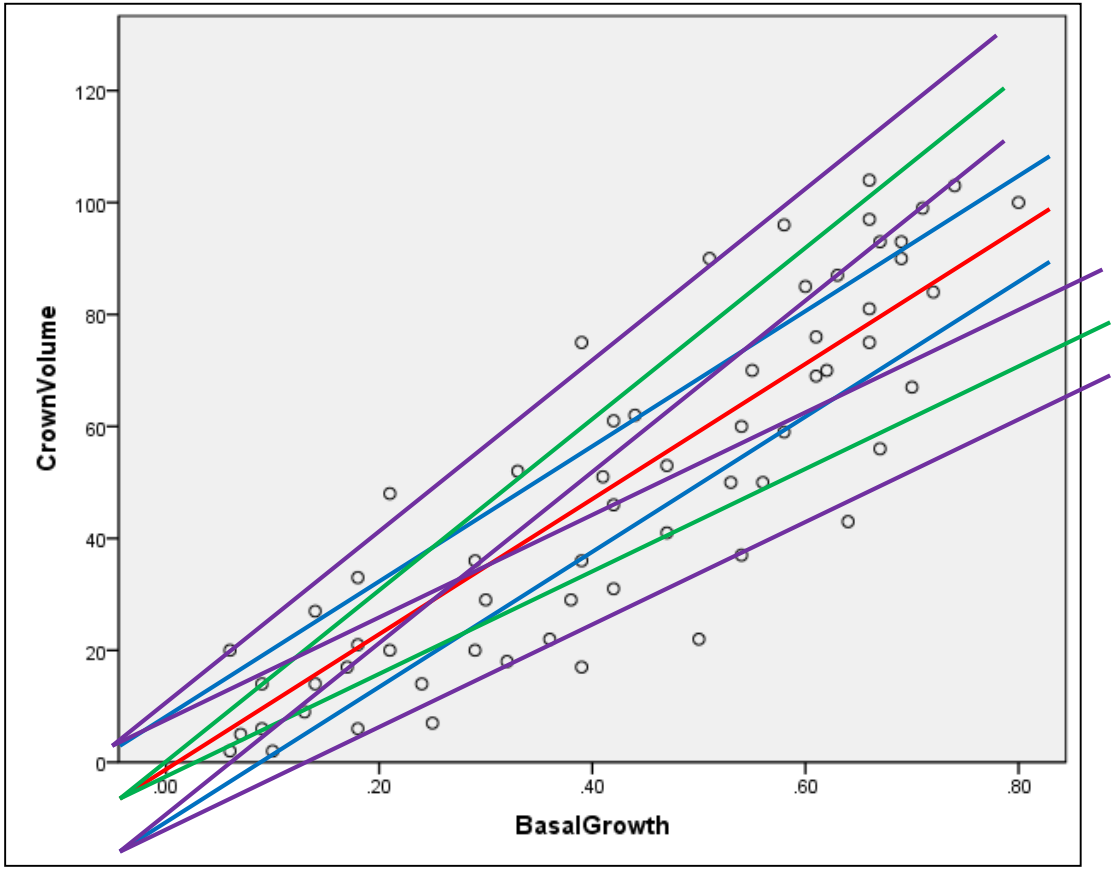
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-5.452	4.435		-1.229	.224	-14.323	3.419
	BasalGrowth	127.273	9.263	.871	13.741	.000	108.745	145.801

a. Dependent Variable: CrownVolume

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Red line: original fitted model
- Blue lines: model with upper and lower confidence intervals for b_0
- Green lines: model with upper and lower confidence intervals for b_1
- Purple lines: model with upper and confidence intervals for both b_0 and b_1



Note: b_0 and b_1 are estimated from the sample so the confidence intervals do not relate to the population

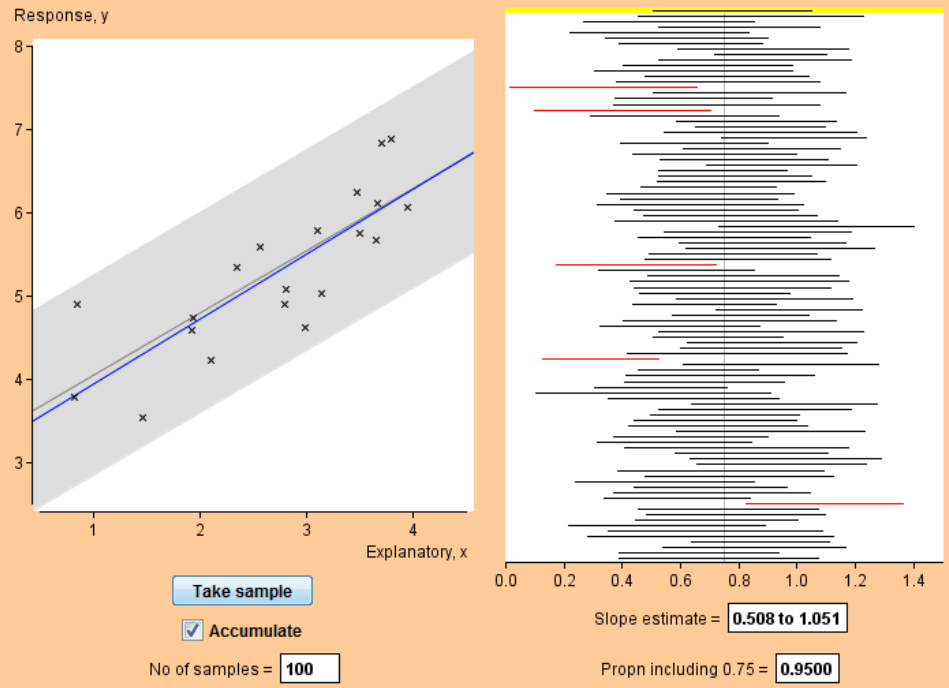
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Simulation of confidence interval of b_1

Simulation

The diagram below shows a sample from a normal linear model in which the true value of β_1 is 0.75. (In real data sets, β_1 is an unknown value but, by simulating data from a situation where it is known, we can examine the accuracy of our estimates.)



On the right, the 95% confidence interval for β_1 based on this data set is displayed. Click **Take sample** a few times to observe the variability in the confidence intervals.

See http://cast.massey.ac.nz/core/index.html?book=general Section 12.2.6

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

The process of regression analysis

Step 1: Get to know your data

- Create a scatter plot, calculate descriptive statistics and look for outliers

Step 2: Formulate a model

- Based on examination of the results of Step 1, hypothesize a model that might explain the data relationships

Step 3: Fit the model to the data

- Examine the regression coefficients

Step 4: Check the fit of the model

- Coming next

Step 5: Report, interpret, and apply the model

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Step 3: Check the model fit

A. Calculate the adjusted R square coefficient
B. Check the significance of the ANOVA model
C. Check the model assumptions, including robustness assumptions

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Check the model fit: A. Adjusted R square

- R square (R^2) is the percentage of variation in y explained by the regression on x
- Adjusted R Square (R^2_{adj}) is the percentage of variation adjusted for the sample size and the number of coefficients in the regression model

BasalGrowth predicts 75.5% of the variation of *CrownVolume*

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.759	.755	15.451

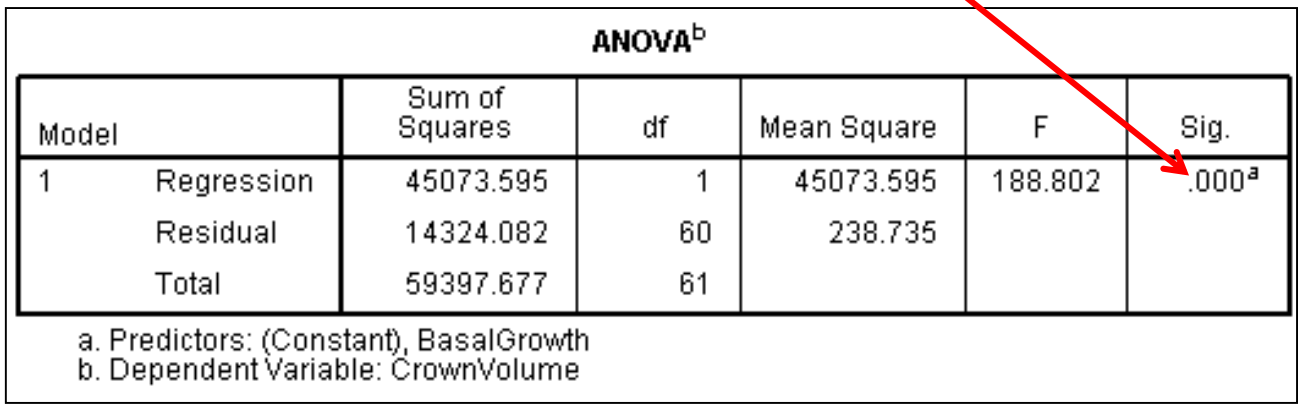
a. Predictors: (Constant), BasalGrowth
b. Dependent Variable: CrownVolume

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Check the model fit: B: Check the ANOVA model for significance

- Generated automatically
- This one is fine ($p < 0.001$)



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45073.595	1	45073.595	188.802	.000[a]
	Residual	14324.082	60	238.735		
	Total	59397.677	61			

a. Predictors: (Constant), BasalGrowth
b. Dependent Variable: CrownVolume

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

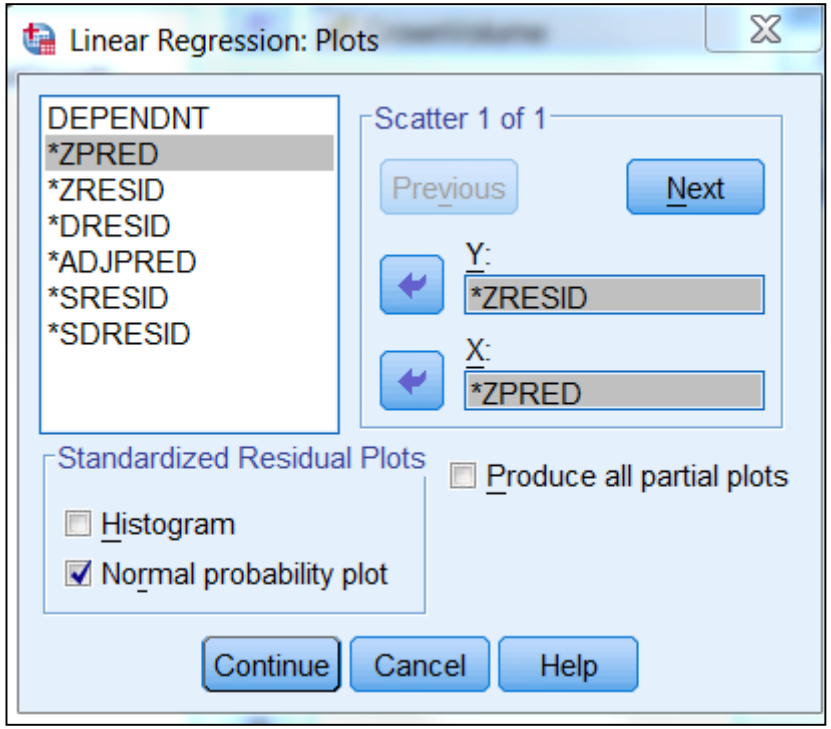
Check the model fit: C. Check the assumptions

- The regression assumptions are checked through analysis of the residuals
- The analysis of residuals is a subjective process of examining:
  - Standardised residuals v. standardised predicted values
  - Normal probability plot
- If the model fails then also check the robustness assumptions (equivalent to normal probability plot fit not being too bad)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Select Analyze > Regression > Linear
- Choose the dependent and independent variable as before
- Select Plots…
- Select *ZRESID for the Y variable (standardised residual)
- Select *ZPRED for the X variable (standardised predicted value)
- Select Normal probability plot

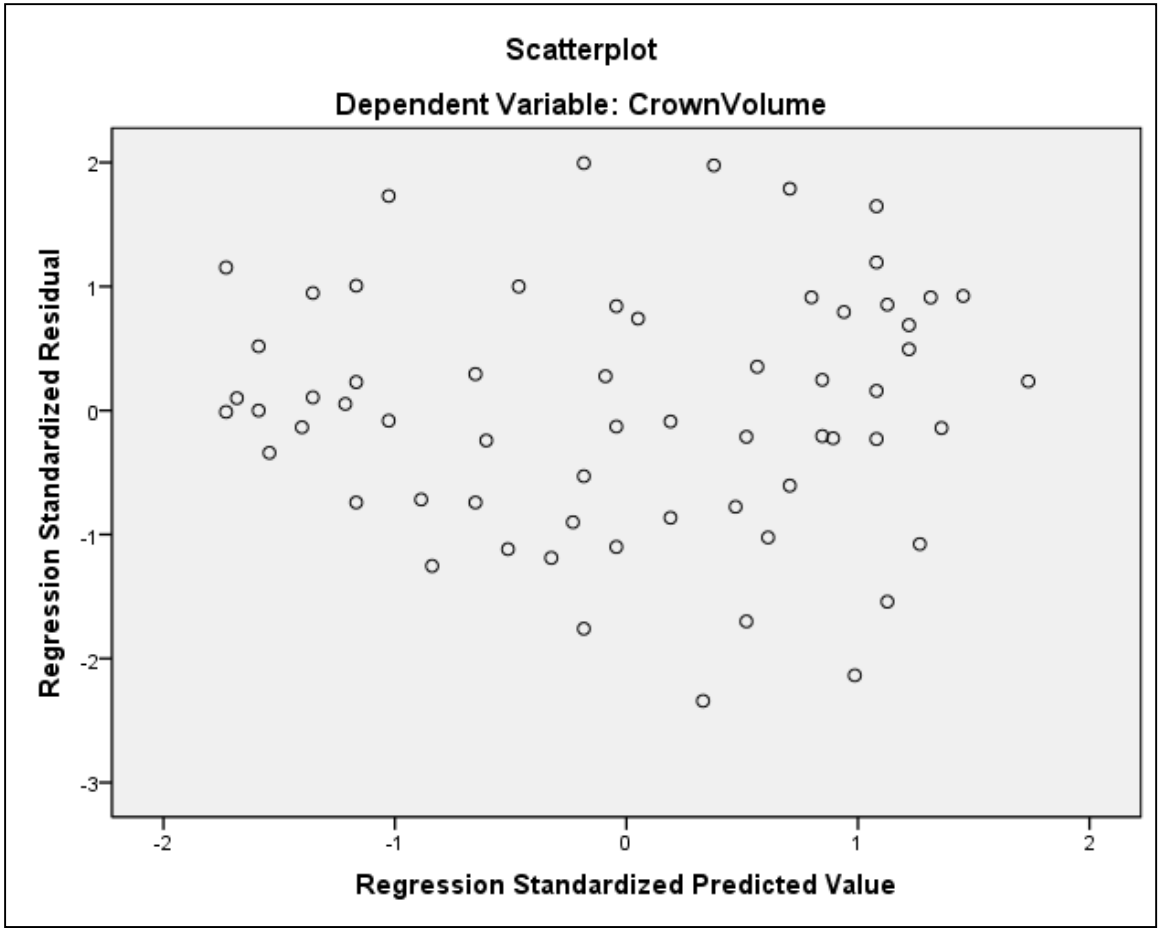


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Standardised residuals v. standardised predicted values

- These should be scattered randomly
- Any discernible pattern (such as a ‘U’ shape) indicates a problem, e.g. not linear
- This one is fine

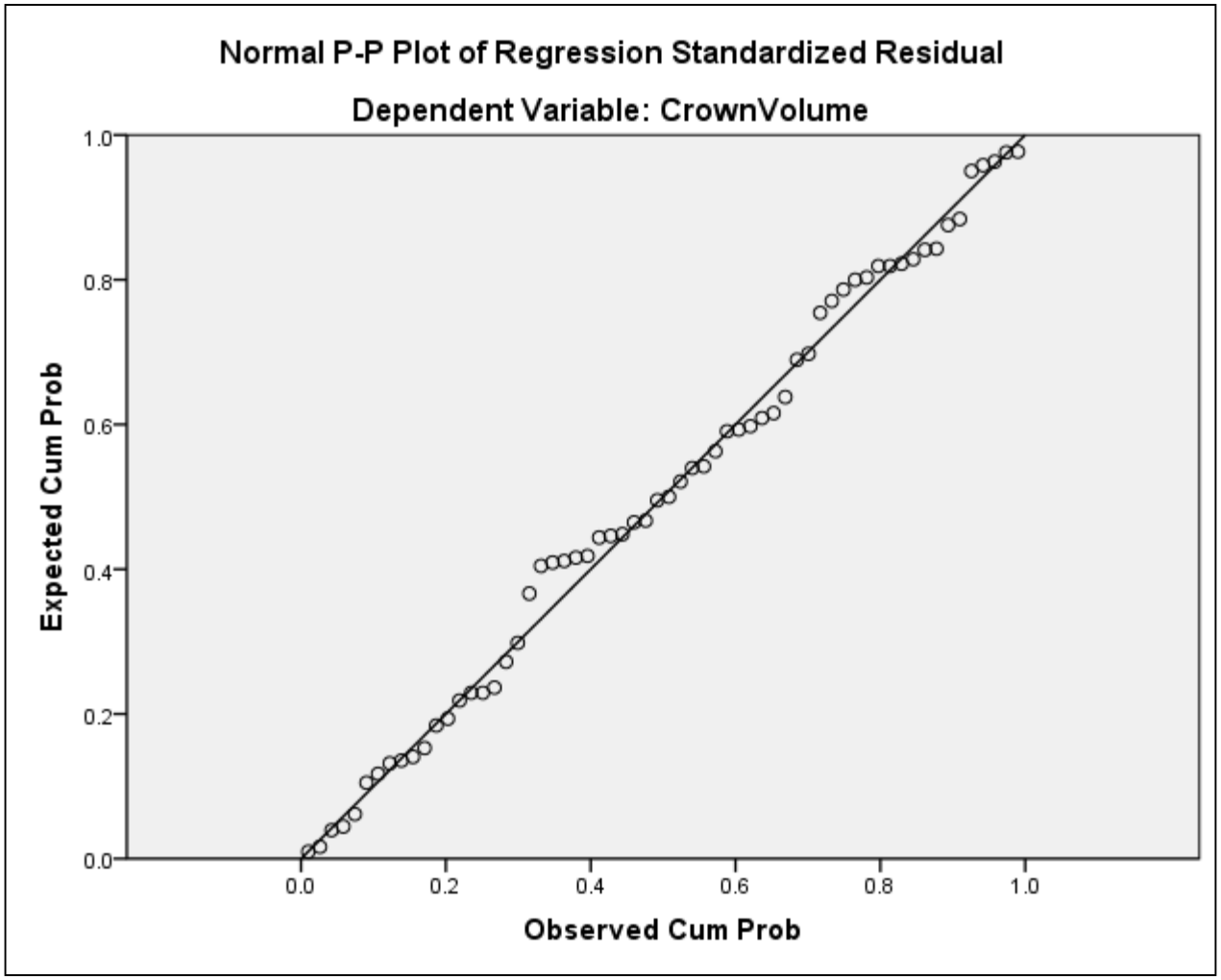


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Normal probability plot

- Normality is indicated by a roughly linear plot
- Any strong systematic curvature suggests some degree of non-normality
- This one is fine



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Robustness exceptions

- Homoscedasticity is mandatory
- Linearity is mandatory
- "Normality is not necessary for the least-squares fitting of the regression model but it is required in general for inference making" (e.g. calculating the p-values and confidence intervals of b_0 and b_1) "only **extreme departures** of the distribution of Y from normality yield spurious results"

  Source: Kleinbaum, D., Kupper L., Muller K. & Nizam, A. (1998) *Applied Regression Analysis and Multivariable Methods*. 3rd ed. Pacific Grove, CA: Duxbury, p. 117

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

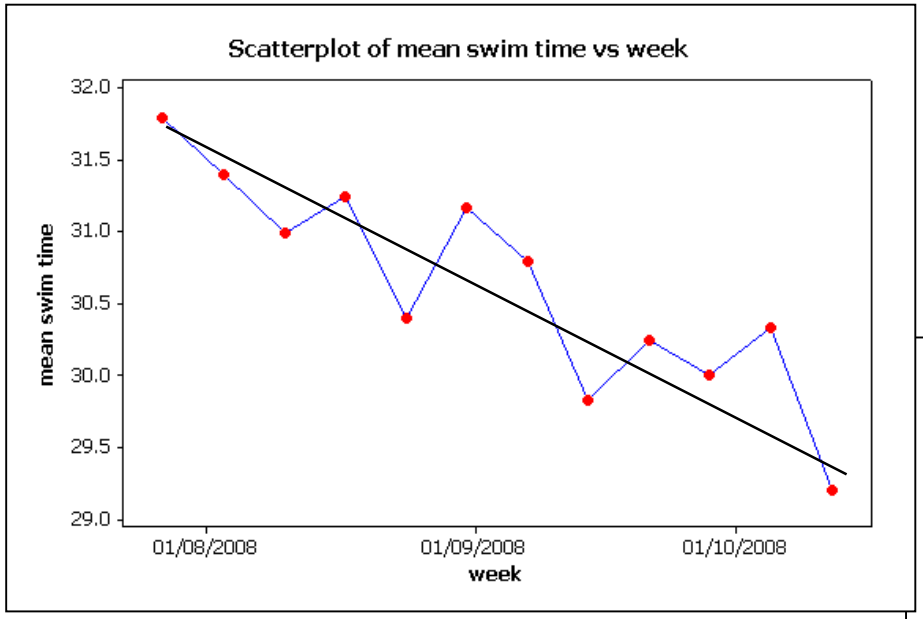
Step 5: Report, Interpret and Apply

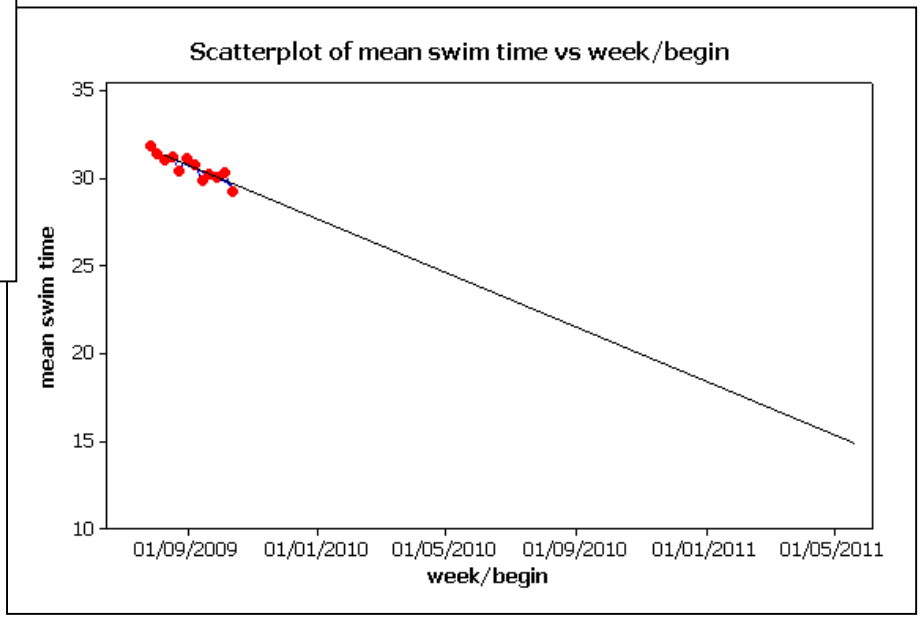
- Report the results of your work in the appropriate context
- Interpret the model
  - Explain the meaning of the coefficients in practical terms
- Apply the model
  - Where appropriate, use the model for prediction

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Don’t extrapolate your data too far away from its range





Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Application to our example

Report:

- From the regression analysis output, basal growth explains 76% of the variation in crown volume increase (ANOVA model significant at 0.001 level)
- The model is:

$$CrownVolume = -5.45 + 127.27 \times BasalGrowth$$

Interpret:

- For every unit change in the growth of the trees there is a 127.3 cubic foot increase in the crown volume.

Apply:

- We may use the model to predict future values of crown growth:

For a *BasalGrowth* of 0.43 square feet,

$CrownVolume = -5.45 + 127.27 \times 0.43 = 49.3$ cubic feet increase

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Regression caveats

- Always plot your data first (Step 1)
- Don’t infer that x "causes" y
- Be cautious about predicting beyond the range of x (extrapolation)
- Beware of outliers
- Examine plots of residuals carefully

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Recap

We have discussed:

- Meaning and computation of Pearson and Spearman correlation coefficients
- Pearson correlation assumptions (including robust exceptions)
- Simple linear models in regression
- The process of simple linear regression analysis
- Simple linear regression assumptions (including robust exceptions)
- Using residuals to check regression assumptions

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Bibliography

Avery, T. & Burkhart, H. (1994) *Forest Measurements*. 5th ed. New York: McGraw-Hill.

Bovas, A. & Ledolter, J. (2006) *Introduction to Regression Modelling*. Belmont, CA: Thomson Brooks/Cole.

Field, A. (2013) *Discovering Statistics using SPSS: (And sex and drugs and rock 'n' roll)*, 4th ed., London: SAGE, Sections 8.1 - 8.4.

statstutor (n.d.) *Pearson Correlation Coefficient resources*. Available at: http://www.statstutor.ac.uk/topics/correlation/pearsons-correlation-coefficient/ [Accessed 8/01/14].

statstutor (n. d.) *Simple Linear Regression resources*. Available at: http://www.statstutor.ac.uk/topics/regression-and-model-building/simple-linear-regression/ [Accessed 8/01/14].

statstutor (n.d.) *Spearmans Correlation Coefficient resources*. Available at: http://www.statstutor.ac.uk/topics/correlation/spearmans-correlation-coefficient/ [Accessed 8/01/14].

Stirling, W. D. (2013) *Welcome to the General CAST e-book*. Available at: http://cast.massey.ac.nz/core/index.html?book=general [Accessed 8/01/14], Sections 3 and 12.

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield